Information Standards Within NCI

Frank Hartel and Sherri de Coronado

As a federal research institution, the National Cancer Institute both funds a large portfolio of external research and conducts scientific research in its own laboratories. It creates, maintains, and analyzes data related to funding grants and contracts; manages a large science portfolio; conducts and shares research; and manages its internal operations. The science supported both internally and externally ranges from the most basic research to large-scale clinical trials conducted by national cooperative clinical trials groups. Now more than ever, with "business to business" interoperation and data sharing among the members of the cancer community, NCI must meet the challenge of managing millions of data points.

Data sharing and process integration are cornerstones of clinical trials processes, and critical to other diverse fields like genetics, molecular biology, and development of human cancer models. The data systems at NCI grew over time; each was designed to meet a specific need or serve a segment of the overall organization. Because aggregation of information from these systems was not a priority, both the vocabulary they used for keywords and coding and the communications and data standards they supported were sub optimized. Only in recent years have the lack of Institute-wide data standards and vocabularies come to be recognized as a strategic problem.

In fiscal year (FY) 1999, NCI Deputy Director for Extramural Research established what would come to be called the Institute Information Systems Advisory Group (IISAG). Recognizing that inconsistent terminology and usage caused serious information search and sharing problems, the IISAG studied how to standardize coding across the Institute. In the resulting report (Gray and Spaholz, 2000), the IISAG recommended that the NCI establish a central repository for NCI vocabulary, centralize coding activities, develop software to support consistent use of vocabulary to store and retrieve data in NCI systems, and establish an Institute-wide oversight group for guidance. The NCI Executive Committee adopted the recommendations and has established an implementation group that began work with the start of FY 2001.

Written from the perspective of vocabulary service developers, this chapter further examines the role of information standards at NCI. It describes the NCI's current use of vocabulary, the governance structure needed to address challenges, and the technology infrastructure that NCI has built to date to support vocabulary services. Included are synopses of major near-term and longer term initiatives and a brief discussion of the importance of intellectual property issues.

Terms versus Concepts: A Vital Distinction

Throughout this chapter we refer to concepts and terms. These are related but not synonymous. It is vital the distinction be kept clearly in mind. We use the word "term" to mean a lexical unit, such as an acronym, word or phrase. We use "concept" to refer to semantic units, that is the meaning of a term. An example makes the distinction clear.

The term “mole” is ambiguous. It can refer to at least three different concepts, a small mammal, a unit of measure, or a skin nevus.

Much of our work has centered on enabling NCI staff to work with terms in automated systems without running afoul of this sort of semantic ambiguity. The vocabulary systems that we build are concept based, because concepts possess attributes like semantic type, which we can use to make their meaning clear.

The Need for Vocabulary Standards

At the NCI, the broad range of research and supporting activities creates many different settings for data creation, maintenance, sharing, and analysis, and therefore many different needs for vocabulary. Since the NCI is first and foremost a research institution, it generates and uses vast quantities of data from basic research laboratories, clinical trials, and epidemiologic studies. Given that NCI supports a large research portfolio, it has an obvious need for information to analyze the science portfolio, from basic research through clinical standard practice. Such information is used to identify gaps and opportunities for new research or initiatives, and for planning, goal setting, and evaluation. These activities can occur only through synthesis of information sources across the Institute, which has been laborious and time-consuming in the past.

Another major use for data and information at NCI is reporting, which ranges from budget to clinical trials summary data. A few examples are:

- reporting expenditures and describing progress in areas of high interest to Congress
- developing ad hoc summaries of research funded in disease areas like breast or prostate cancer, or science areas like potential anti-angiogenesis treatments
- summarizing gender and minority accruals to clinical trials
- reporting to the FDA adverse events that occur during clinical trials
- answering Freedom of Information Requests about specific grants or contracts
- disseminating cancer incidence and mortality statistics

The sources of information for reporting are generally databases: e.g. grant databases, clinical trial databases, databases that collect incidence and mortality data from cancer registries, and science databases on such topics as genetics. Most of the data in them are “coded” with vocabulary terms to enable retrieval for reporting.

The NCI also uses data and information for its publications, many of which appear on the Web. A broad range of information is contained in these sources. For instance:

- Clinical trials information includes eligibility criteria, protocols, treatment statements, and other related documents.
- The NCI website contains a variety of information about research programs, funding opportunities, cancer trials, and other resources for scientists and the public.
- Annual research and budget documents that are frequently accessed by the public.

Much of the data for publication originates from the same databases described above, plus several others. Information made available through the Internet is ripe for "coding"[1] or indexing to facilitate retrieval, and NCI staff are recognizing the potential value in providing links to definitions of the cancer concepts in these documents, as well as broader and more specific terms users may not understand.

Several issues emerge from this need to create, use, synthesize, and share information—issues that drive the need for a shared vocabulary.

- **Consistency and Completeness.** For financial reporting, as well as for other published data and statistics accessible by the public or scientists, consistency and completeness of retrievals is paramount.
- **Data Sharing.** Science does not operate within well-defined boundaries; it is multi-disciplinary and advances through the sharing and synthesis of multiple sources of data. For example, data from genetic, pathology, and clinical trials databases need to be combined and shared to clarify the molecular basis of cancer.
- **Storing and Communicating Information.** Where and how to store information that will be used for analysis is also an issue. Should users create data in their own databases using a common vocabulary, should they enter their data into large shared repositories, or should they do post-facto mapping of their data to the standards of a "group" database?

Nearly every time a database is developed at NCI, the developer needs to adopt or create a "coding" system that will represent the content of the database and facilitate retrieval and analysis. The database owner would need to "interpret" the results of a search and retrieval report for a data consumer. Over the years, databases have adopted their own idiosyncratic ways of representing content; however, as the need to share and synthesize data across databases has grown exponentially, it has become clear that a shared (standard) vocabulary is needed. The science is driving the need for vocabulary standards that will ensure consistency and completeness of retrieval and enable data from multiple retrievals to be combined in a meaningful way.

Vocabulary Challenges at NCI

The challenges facing NCI with respect to controlled vocabulary and related standards reflect the diverse nature of the Institute's activities. Because of its clinical trial activities, NCI has many of the vocabulary and data interchange requirements familiar to hospitals and clinicians. NCI is engaged in basic biological research and applied biological and biomedical research, and these activities generate extensive vocabulary and data interchange requirements. Administrative activities, too, have their own distinct vocabulary and data interchange requirements.

Even within the context of clinical trials however, there are significant differences between the NCI environment and that of typical clinical medicine. For example, reimbursement—a major focus of reporting and data interchange in the typical hospital

[1] Assigning alphanumeric strings to a data base record to facilitate retrieval.

setting—has been of small consequence for NCI. On the other hand, regulatory reporting and correlation of detailed, voluminous data from multiple institutions looms large in NCI clinical trials. Commercially available medical information systems meet some, but not all NCI clinical needs, and no single standard vocabulary or set of vocabularies is adequate.

Biology and biomedicine present the greatest vocabulary challenges. These areas generate a high volume of new knowledge, and the coverage of research concepts in standard vocabularies is relatively shallow. Mechanisms are needed to rapidly model new concepts in controlled vocabulary as they surface in sources like Entrez, (http://www.ncbi.nlm.nih.gov/entrez/) and to provide the vocabulary to NCI for operational use.

To meet its administrative needs, NCI has developed a highly specialized vocabulary. The administrative vocabulary is used principally for Congressional and management reporting, as well as financial accounting and oversight of grants and contracts. NCI administrative activities require a tightly controlled vocabulary, so that reports and analyses are consistent and comprehensive.

These three vocabulary areas—clinical, biological, and administrative—are interdependent. As new concepts enter in the research vocabulary, they quickly begin to be used in the NCI clinical trials and grants administration applications. This requires that the vertical and the horizontal relationships among NCI concepts are simultaneously valid in the ontological sense and navigable by members of many disciplines.

The Problem of Hierarchy

Ideally, in formal hierarchies, vertical relationships are “is_a relationships” in which everything true of the “parent” is true of the “child.” In practice organizing principles such as “part_of” are needed, for example in anatomy hierarchies. Not with standing the organizing principle, when vocabulary is used as keywords or retrieval codes in a database, the hierarchical relationship of the concepts must be valid.

For instance, queries against NCI Enterprise Databases will depend on vocabulary servers for query “explosion” and aggregation. When a query is exploded, the database user specifies search term. The database passes them to the vocabulary system. The vocabulary system identifies the retrieval concept referred to by the search term, and then walks the vocabulary tree downward starting with the specified concept, and returns all the “children,” or results, to the user’s database. The user’s database system then constructs a query consisting of the original concept and all its children. Conversely, in an aggregation, the user starts with a concept, is shown all its parents, and selects among the broader concepts for one at the desired level of generality. If NCI Enterprise Systems are to generate search results that are consistent and comprehensive, NCI concept trees must correctly embody the parent-child relationships as the communities within NCI have defined them.

The Problem of Semantic Relationships

Horizontal relationships among concepts specify how a concept is semantically related to another. Given the need for navigability by a diverse population of users, the NCI

vocabulary systems use semantic relationships among concepts to provide links between the clinical, research, and administrative areas. For example, the administrative vocabulary possesses a number of reporting categories that are generally broad and may have no valid clinical or scientific meaning. However, they do have important social, political, or other meaning, and are related in some sense to sets of scientific or clinical concepts. “Head and neck” may not be a meaningful anatomical term, but it is recognizable as a token for a set of anatomical sites where specific oncology diseases occur. In this case, no valid “is_a” relationships may exist, but valid semantic relationships of some sort most likely do. By modeling the semantic relationship explicitly (Campbell, 1997), the NCI vocabulary system can help the clinician understand if a specific disease site is considered “head and neck” for reporting purposes.

There is danger that the burden of modeling and maintaining such semantic relationships will become overwhelming. For this reason, modeling must be limited to relationships of clear value to NCI operations.

The Problem of Representing Molecular Biology and Medicine

Arguably the greatest challenge facing NCI’s vocabulary and standards work is dealing with the explosion of molecular biology and its accelerating impact on cancer medicine. Cytogenetic notation has begun to appear in standard vocabularies like ICD-0-3 (Harris, et al.), as prognostic indicators and as diagnostic criteria. A standard way of representing this and other molecular biology information in the context of controlled medical vocabulary must be developed. NCI may be forced to address this issue for cancer-related vocabulary sooner than others (Klausner, 1999).

More and more novel therapeutic interventions are being developed, like ONYX-015, a genetically engineered adenovirus that, in preclinical and clinical studies, has been shown to replicate in and kill tumor cells deficient in p53 tumor suppressor gene activity (Makower, et al.). As the mechanisms of action for these new interventions are understood, and as they are proven useful and enter the clinic, NCI will have to develop principles governing how to place them in a hierarchy (or “tree” them) and semantically relate them. A major goal of these principles will be to avoid continual rework to the fundamental vocabulary modeling as scientific and clinical details emerge.

Collaborations with the Cancer Community

NCI currently has a large number of home-grown vocabularies in use; however, standard vocabularies are used in several areas of the Institute, most notably in the clinical area. In these areas, NCI participates in vocabulary development with the cancer community.

The Surveillance, Epidemiology and End-Results Program (SEER) program, for example, uses the ICD-O-2 vocabulary for collecting and coding data on incidence and survival from cancer data registries around the country. (See: http://www-seer.ims.nci.nih.gov/ .) The ICD-O-2 was developed by the World Health Organization (WHO) with substantial NCI participation. NCI also participated in the development of the ICD-O-3 that will be released by the WHO in early 2001. Other areas of NCI that have not previously used the ICD-O are awaiting the release of the new version, since it

is the first vocabulary to include relevant cytopathology information as part of the disease classification.

The Cancer Therapy and Evaluation Program (CTEP), which manages a large portfolio of funded clinical trials, is required to report adverse event information to the FDA. The vocabulary used to code these adverse events is the Medical Dictionary for Regulatory Activities (MedDRA) vocabulary, formerly International Medical Terminology (IMT). The researchers responsible for these clinical trials report the adverse events to the NCI using this vocabulary, and NCI transmits the data to FDA. MedDRA is another internationally developed vocabulary in which NCI participated, contributing heavily to a major part of MedDRA, the Neoplasms, Benign and Malignant System Organ Class (SOC). (See: http://www.msso.org/default2.htm)

The Physician Data Query System (PDQ), a part of the Office of Cancer Communications (formerly the International Cancer Information Center), incorporates a vocabulary used extensively in the Cancer Community and by the public to retrieve information from that system. While not a nationally or internationally developed vocabulary, it has been submitted to the UMLS for several years, and is therefore made available to the community to use as it wishes. The vocabulary is currently undergoing extensive revision and review, and the new version will be used by the PDQ database and other CancerNet databases and submitted to the UMLS in the future. The PDQ staff currently also provides a glossary of lay cancer terms for use with PDQ and NCI web documents. (See: http://cancernet.nci.nih.gov/pdq.html)

Development of Basic Science Vocabulary

While the clinical side of the Institute has a history of working with the community in developing and using standard vocabularies, basic cancer researchers and science managers have only recently begun to interact with the cancer and biomedical communities in vocabulary development and use. In the past, the basic science vocabularies used by NCI have been mostly small and home-grown, sufficient for program managers to answer individual questions about the relatively small portfolios they manage, and for a central indexing organization to produce information for NCI budget reports. Although this satisfied the need to answer questions like “How much did the Institute spend in a particular disease area or special interest area in a particular year?,” it did not help answer questions about the science itself or facilitate collaboration with other agencies or interdisciplinary research programs.

The need to answer science questions across multiple data sources requires standard vocabularies that are detailed enough to describe the science precisely, especially since the distinction has blurred between information about the science and the science data itself. Consequently, the Institute has engaged in several basic science vocabulary related activities that address both the science and the science information aspects.

One such activity is NCI’s ongoing effort to “model” the vocabulary for cancer related genes and proteins and provide that information to the UMLS as it becomes available. This will benefit the entire biomedical community and spur further development. Another activity, the Mouse Models of Human Cancer Consortium (MMHCC), is beginning to build a vocabulary that will help clarify the differences and

similarities in mouse and human cancers so that better models can be built. A completely different type of activity is the NCI's Common Scientific Outline of cancer research, developed jointly with the Defense Department to enable exchange of information about the NCI and Defense Department cancer research portfolios.

An NCI database that contains spectral karyotyping and comparative genomic hybridation data for the high-profile Cancer Chromosome Aberration Project is being designed for international use. The scientists and database designers, working with an external group of scientists who will also contribute data to the database, have decided to use the ICD-O-3 as the standard for assigning topography (site) and morphology (e.g. cell type) coding to their sample data. This will be one of the enablers for combining data from many different research projects. By providing access to the NCI Enterprise Vocabulary System (discussed later) to potential users and contributors of the data, those who choose to use a different standard vocabulary like SNOMED International in their own systems can map to the ICD-O-3 vocabulary for placing data in the "shared" database.

A relatively new effort to create and maintain standard vocabularies for NCI clinical trials is discussed in another chapter. Called the Common Data Elements project, it heavily involves the cancer community in developing vocabulary relevant to particular clinical forms like case reports, along with values for these data elements in the context of the appropriate form. Other Common Data Elements projects are expanding the scope of the vocabulary under development to pathology, radiology, epidemiology, and other information relevant to the clinical trials.

Increased Recognition of the Need for Vocabulary Standards

Over the past year, a wide variety of players at NCI—scientists, administrators, planners, systems developers, program managers—have recognized that vocabulary standards are crucial to the continuing progress of cancer research. It is a strategic issue, since it provides the only way to meaningfully share data, and it will be vital to linking the basic and clinical research that will lead to new treatments.

Willingness to participate and engage in discussions and planning has increased markedly. Top NCI leadership has recognized the need for an Institute-wide vocabulary review board. We are of the opinion there ought to be an advisory component staffed with leading experts from outside NCI to provide external input into NCI vocabulary activities. Further, many NCI components have realized that it is easy to create vocabulary, but difficult to maintain it. Thus, a major implication of the recognized needs will be additional use of standard vocabularies that are maintained by the scientific and medical communities.

Emerging Trends

Several trends within and outside of NCI have underscored the importance of vocabulary development and standards. These trends provide a focus to NCI's initiatives.

Migration from Isolated Systems to Integration

Several years ago, the NCI Director and other senior managers envisioned a system that would answer questions about the science NCI supports, clarify the gaps and opportunities in the portfolio, and synthesize information from many different systems that historically have been isolated. This vision has driven NCI to move towards integration. As the size of the NCI program, the pace of discovery, and operational requirements for greater integration all increased, NCI undertook major investments to create enterprise systems that shared a data model and adhered to defined standards. These systems aimed to help NCI develop integrated applications supporting all critical internal activities and share data with business partners in government and the research sector. Many applications that access this database are integrated in an internal NCI website called the NOW (NCI Online Workplace).

The availability of enterprise systems means that given a robust vocabulary and business rules about how to do "coding," the NCI now has the means to store the various sets of coding that exist for a given grant or contract. Coding from various programs can be combined in a single database to enable single queries about a topic across the portfolio using a common language. This moves the Institute considerably closer to its vision.

An exemplar of this integration model is the Science Place, a knowledge management application that may be used to support the work of the Mouse Models of Human Cancer Consortium or other specific NCI activities. This application provides a place for scientists, analysts, and managers to retrieve information from many sources, organize that information according to their own preferences, and share information about topics in which they are experts. The Science Place depends on the rich synonymy provided by the NCI Enterprise Vocabulary System, which enables search, retrieval, and organization of information from bibliographic databases, genetics databases, the Internet, its internal database, and NCI science data warehouse.

The Coming of Governance

Although NCI systems like the Science Place can retrieve comparable information in several databases even if the information is coded with different keywords, it is better to avoid inconsistent vocabulary to begin with. Avoiding them is a major goal of governance. Governance is the best way to sidestep the "Tower of Babel" effect in databases.

Recently, the Long Range Planning Committee (LRPC) advised the NCI that establishing new processes to speed development, approval, conduct, and reporting of clinical trials would require considerable investment in governance (Chute and Langlotz, 2000). Because of NCI's central role in funding and conducting cancer-related research and in translating findings into cancer care, decisions the NCI makes about vocabulary, messaging, and metadata standards affect the operations of others. It is for this reason that we advocate that NCI establish an external advisory group. We feel the NCI's governance procedures must reflect the needs of both its business partners and the whole cancer community.

The Long Range Planning Committee recommended that NCI establish a standing advisory group, composed of leaders from across the cancer community and various standards development organizations. The group would provide strategic advice to NCI

regarding vocabulary and health-related standards and critique NCI's decisions and activities in these areas. The Committee also recommended that NCI become the focal point for representing the cancer community in the standards development organizations.

Laying the Groundwork: NCI EVS

Well before the consensus formed that vocabulary standards and governance were needed, NCI began to lay the infrastructure to support them. The NCI Enterprise Vocabulary System, or NCI EVS, provides a variety of vocabulary-related services to NCI. Among them is the NCI Metathesaurus, which is a core infrastructure component. Created to neutralize the Tower of Babel effect that had grown among legacy NCI systems, The NCI Metathesaurus is a technology-based stopgap that has enabled NCI, for the first time, to:

- identify gaps and inconsistencies among the informal terminologies being used by the Institute
- provide a central point for maintenance of these terminologies
- provide a single resource from which all NCI systems could obtain vocabulary
- place the NCI terminology into trees and map it to standard vocabularies

The NCI EVS infrastructure consists of technology and human resources. The technology includes server hardware, database, editing and management software, and applications programming interface software. The human resources include operations staff to care for the servers, vocabulary databases and software, curators who edit and maintain the content of the vocabulary databases, and applications support staff who interface the NCI EVS to other NCI systems. Figure 14.1 provides an overview of NCI EVS processes and components.

INSERT FIGURE 1 HERE ****see attached files as revised****

Licensed Software

NCI has attempted to construct the NCI EVS from commercial products that appear to be emerging de facto or formal standards. The Apelon Incorporated Architect™, Authority™, Concept-based Retrieval™ System and Metaphrase™ products are being used operationally, and the Mayo Vocabulary System, the Food and Drug Administration's Autocoder and Apelon TDE™ (Terminology Development Environment) and DTS Metaphrase™ (Distributed Terminology System) products are being actively evaluated. These products are attractive to NCI because, except for Autocoder, they support "open" Java APIs. In the case of the Architect™, TDE™ and Mayo products, they are also compatible with the description logic (Campbell, 1997) vocabulary representation that is becoming widely used in private sector entities like the College of American Pathologists (Hochhalter, 2000;College of American Pathologists, (#1)) and government initiatives like the Government Computerized Patient Record consortia (Brown, 2000), the National Library of Medicine government-wide

SNOMED/RT acquisition, and National Health Service of the U.K. (College of American Pathologys, 2000, (#2))

Vocabulary Services to NCI

Two distinct vocabulary service offerings are needed to meet NCI's needs. The NCI Thesaurus is a description logic based vocabulary that contains only NCI terminology. The NCI Metathesaurus contains many of the sources contained in the UMLS Metathesaurus, plus the NCI vocabulary. Tailored to the needs of NCI database systems, the NCI Thesaurus ensures formally correct concept modeling, which in turn provides reliable navigation among NCI concepts and correct explosion and aggregation of search terms. These properties are vital to consistent and comprehensive retrieval from NCI databases. The NCI Metathesaurus provides rich synonymy, English language definitions, and mappings between NCI terminology and sources like SNOMED, ICD, and MeSH, which are used within NCI or by NCI business partners. The NCI Metathesaurus is especially useful to NCI Webmasters for indexing documents and helping users navigate to the biomedical concepts in which they are interested.

EVS Editing, Review and Change Management

In FY 1999, NCI decided to employ contractors to edit the NCI EVS content. Review would be provided by NCI staff and by outside reviewers. NCI EVS editors use the Authority™ and Architect™ editing environments for content creation, and new releases of the NCI Thesaurus and NCI Metathesaurus are periodically produced and made available for testing. These activities are governed by a configuration management and change management plan. Change requests and dispositions and configuration control activities are tracked and analyzed. These practices bring the NCI EVS close to compliance with the ASTM E2087 Standard Specification of Quality Indicators for Controlled Health Vocabularies. Plans to achieve full compliance with this standard and a variety of others are discussed elsewhere in this chapter.

Vocabulary Update Processes

The NCI Thesaurus is continually updated. Each month, new NCI Thesaurus releases are issued for use by NCI databases and other systems and imported into the NCI Metathesaurus using the Authority tool. These minor monthly releases of the NCI Metathesaurus ensure that the rapid evolution in biomedical terminology, especially in the areas of cancer genetics, cell and molecular biology, and other fast moving research areas, is available to NCI Metathesaurus users. Once a year, the NCI issues a major release of the NCI Metathesaurus, which keeps the NCI Metathesaurus current with new releases of the National Library of Medicine's UMLS Metathesaurus.

Because NCI concepts "inherit" synonyms and other relationships from Metathesaurus sources, "false" synonymy, definitions, and other relationships to NCI concepts occasionally crop up. NCI editors use the Authority tool to delete many of these, but some cannot be eliminated without editing non-NCI sources. Remaining false relationships are the unavoidable price paid for the rich synonymy and other benefits of

embedding the NCI Thesaurus in the Metathesaurus-like environment of the Metaphrase Metathesaurus™. Because of this relative lack of control, most configuration management and quality review effort is directed at the NCI Thesaurus, not the NCI Metathesaurus.

Near Term Initiatives

To realize the benefits of internal information sharing and business to business operations that an investment in infrastructure can provide, the NCI must implement the governance structure alluded to earlier in the chapter. In 2001, much of the governance structure envisioned by the Long Range Planning Committee and the IISAG Coding Report should be in place.

External Experts and Internal Stakeholders

The NCI's vocabulary initiatives must deliver appropriate products and well-managed services to clients within the NCI and, where appropriate, to those in the broader cancer community. As recommended by the LRPC, an external advisory group will be needed; as recommended by the IISAG Coding Committee Report, an internal oversight group will be needed. We advocate that these groups be called the NCI Vocabulary and Standards Advisory Committee and the NCI Vocabulary Executive Group. Their combined goals ought to be to:

- guide the NCI's efforts in vocabulary use and development
- set expectations for how vocabulary will be used within the NCI and how the vocabulary resources can assist the cancer community
- oversee the provision of quality vocabulary services to the NCI and to the cancer community
- maximize the coherence and interoperability between NCI's vocabulary efforts and those taking place outside the Institute

Figure 14.2 shows how we envision the relationship of these two groups to each other and to the components that have operational responsibility for vocabulary service and implementation of information-related standards.

INSERT FIGURE 2 HERE

External Advice and Communication

The NCI Vocabulary and Standards Advisory Committee could be created under the aegis of the NCI Director's Advisory Group. The individual roles of the NCI Vocabulary and Standards Advisory Committee should be to:

- maintain a high level of awareness regarding the directions of vocabulary-related products, services, and standards in external communities of interest to the NCI
- convey this information to the NCI Vocabulary Executive Group for further dissemination within the NCI

- maintain awareness of NCI initiatives and needs regarding vocabularies and related standards and ensure that they are clearly presented to relevant parties outside the NCI
- provide advice to the NCI Director
- serve as liaison to the cancer community, the standards-setting community, and the publishers of vocabularies, ensuring an effective bi-directional flow of information, concerns, and plans

Members of this advisory committee should be prominent and influential within the communities in which they are active. They should be chosen to represent relevant standards development organizations, cancer researchers and clinicians, advocacy groups, industry, and other influential stakeholders in the cancer community. Advisory committee meetings should take place two or three times per year.

Internal Oversight and Direction

The responsibilities of the NCI Vocabulary Executive Group should consist of an overall mission and three more specialized roles. Overall, the group should:

- ensure that NCI's vocabulary efforts maintain maximum coherence with the direction of industry technology, methods, content and standards
- ensure that software and vocabularies acquired and developed by NCI will have lasting value by maximizing interoperability with those outside the Institute
- ensure that NCI develops clear and persuasive presentations of its needs for vocabularies and related products, services, and standards
- convey these needs through participation in external activities and through the efforts of the NCI Vocabulary and Standards Advisory Committee

The three more specialized roles concern vocabulary, operational issues, and technology-oriented factors. For vocabulary, the Executive Group's responsibilities should be to set goals, guidelines and procedures for the use of vocabulary at NCI, select commercial vocabularies for integration with the NCI/EVS, and oversee editing of the NCI vocabulary. Responsibilities for operational issues include monitoring delivery of vocabulary-related services, acting as advocate for NCI organizations where service falls short of needs, and serving as the NCI/EVS Configuration Control Board. Finally, for technology-oriented factors, the group should act as a liaison with publishers of vocabularies and represent the NCI on selected standards groups.

The NCI Director or Deputy Director should charter this Executive Group, the membership of which should be NCI staff members. Senior staff representing the entire NCI organization at the Division level should be included, and other NCI staff will be nominated as appropriate to supply specialized knowledge or to meet specific responsibilities. The group should meet quarterly.

In connection with vocabulary services and information standards related to medical or biological requirements, NCI's Center for Bioinformatics should have a range of responsibilities:

- assessment and selection of commercial technology and methods for vocabulary services
- development and testing of new technology for vocabulary services
- advice and assistance in using vocabulary resources
- vocabulary editing and other operational responsibilities
- operation and maintenance of vocabulary-related software
- representation of the NCI on IT-related standards groups

The Center for Bioinformatics will depend on the NCI Office of Information Systems and Computer Systems for operations of servers.

Standards Relevant to Cancer

We envision the NCI becoming the focal point of the cancer community with respect to information standards and vocabulary. Within this vision, more fully described in the Long Range Plan, the NCI Vocabulary and Standards Advisory Committee would serve as an ongoing forum for dialog between the standards development community and the cancer community. Determining which standards are relevant to cancer would be one result of this dialogue. Table 14.1 contains a summary of several standards with clear relevance. Others will emerge in time.

NCI has been an informal—or, in several cases, formal—participant in some Standards Development Organization (SDO) activities, and it should become a formal participant in all SDO activities judged relevant to its information needs. Through the Vocabulary and Standards Advisory Committee and other outreach activities, NCI representatives should become aware of the needs and opinions of the broader cancer community. They should represent NCI and the larger community to the SDO and inform the NCI and the community of important plans and decisions.

During 2001, the NCI will likely commence formal membership in each of the SDO activities listed in Table 14.1. The Center for Bioinformatics is developing a website devoted to standards and vocabulary, which will be used to foster communication among the NCI representatives to the SDO and between the cancer community and the NCI representatives.

INSERT TABLE 1 HERE

The NCI representatives to SDO activities will largely be drawn from the NCI Center for Bioinformatics, since these activities principally involve information technology.

Table 14.2 lists vocabulary products known to be relevant to NCI. During 2001, as with SDO participation, NCI should seek formal representation on editorial boards or similar entities for each of these vocabularies. Since the focus will be on issues of biomedical terminology, most NCI liaisons to these vocabulary developers should be members of the NCI Vocabulary Executive Group.

INSERT TABLE 2 HERE

Each vocabulary in Table 14.2, except SNOMED, is needed for NCI operations. SNOMED is included because of ongoing efforts by a consortium of federal agencies to license SNOMED/RT for use across the Federal government. In anticipation of this license, the NCI is investigating formal collaborative arrangements to facilitate rapid migration of new cancer-related concepts, especially cytogenetic and molecular concepts, from NCI Thesaurus to SNOMED/RT. The technical infrastructure to do this is largely in place, since the NCI Thesaurus is being developed using tools and description logic similar to those used by The College of American Pathologists (College of American Pathologists, 2000) in developing SNOMED/RT (http://snowmed.org).

The NCI Vocabulary Executive Group should make decisions about which standard vocabularies to license and how NCI should use them. Direction of NCI Thesaurus development and usage within NCI should also fall to the Executive Board. As mentioned previously, an extensive configuration management process has been developed. It will enable the Executive Board to develop policies and practices that will ensure that all users depending on the NCI Thesaurus are made aware of changes to its content, especially changes to the concept hierarchy that could directly affect query performance of databases.

The NCI Thesaurus and NCI Metathesaurus are largely compliant with ASTM E2087 quality indicators, and work continues to make both fully compliant in 2001. Efforts to register the NCI Common Data Elements with HISB USHIK are ongoing, and the NCI will begin making its systems HL7 compliant, especially any enterprise systems that must interact with clinical grantees and other clinical partners.

The NCI Vocabulary Executive Group should work with the NCI Vocabulary and Standards Advisory Committee to determine priorities for achieving standards compliance within NCI. They should also consider to establishing measures for assisting business partners or other members of the cancer community to achieve compliance.

Business Case for Coding

At a minimum, NCI has two business goals for its use of controlled vocabulary and standards: operational efficiency and improved scientific productivity. Operational efficiencies between NCI and external entities—the ability both to share information and to integrate systems—will result in cost savings. For example, clinical trials management procedures using the CDE data set and business-to-business techniques promise to save both NCI and trialists money. More importantly, it will provide for much better clinical management, due, for example, to adverse event information being rapidly disseminated to all relevant protocol directors.

Within the NCI, controlled vocabulary and information standards will enhance research productivity. It will be much easier to find and interpret information relevant to a scientific issue from the multiple systems that contain grant, contract, internal project, and other scientific information. This will benefit both the NCI researcher and NCI research management. It will also greatly reduce the burden of generating routine and ad hoc reports for Congress, the Department, oversight and advocacy organizations, and the press.

Uniform Coding and Keyword Practices

The NCI Vocabulary Executive Group should determine how the licensed and NCI-specific vocabulary is used, both for coding and key wording artifacts in data systems and as aids for search and retrieval in websites and other information resources. The IISAG Coding Committee's recommendation that coding and keyword assignment be done by a dedicated, centralized group of experts will be the point of departure. Whatever the Executive Group decides to do with respect to code and keyword assignment, uniformity of practice across the NCI will be vital.

The criteria that will be used to determine if the NCI's code and keyword strategy is satisfactory should be empirical, and results-based vocabulary assessment should be adopted. This would depart from historical practice at NCI, where code and keyword practices were often driven by the limitations of technology or of available staff or organizational culture. If vocabulary or code/keyword decisions do not result in measurable improvement in the comprehensiveness or consistency of NCI information systems, they should be considered unsuccessful.

Coherence Across NCI Involvement with SDO Activities

Because the NCI is so large and diverse, it is not surprising that many of its components are involved with various vocabulary developers and SDO activities. The NCI Vocabulary Executive Group must ensure that the Institute presents a consistent face across these interactions. In many cases, the NCI is engaged in informal interactions with clearly important vocabulary and standards groups, and the Vocabulary Executive Group should formalize these, with NCI becoming a formal member of important development efforts. These steps are the *sine qua non* of any NCI effort to become the focal point for cancer community interaction with the SDO and vocabulary developers.

Investment in Standards as Community Resource

To become formally involved in development efforts and establish reliable means to ensure that the NCI and the cancer community agree about standards and vocabulary, the NCI will have to undertake significant financial investment. More will be needed if the cancer community is to benefit fully from NCI investments in standards development and compliance.

The ongoing effort to license SNOMED/RT across the government aims to cover use of the vocabulary by external entities in their interactions with government agencies. In effect, such a SNOMED/RT license would shelter these entities from the cost of licensing vocabulary so they can do business with the government. This license procurement, then, can be seen as the government "buying down" the cost of adopting a standard for business-to-business communications. The NCI may need to establish other ways of encouraging business partners to consider early adoption of standards-compliant systems or vocabulary. In the absence of such inducement, the pace of replacing old, non-compliant technology may well be too slow to meet the country's need to translate rapid cancer research advances into prevention and care improvement.

As mentioned previously, NCI is discussing a cooperative research and development agreement with the College of American Pathologists to help migrate new terminology from the NCI Thesaurus to SNOMED/RT. The Institute is also contributing portions of the NCI Thesaurus to National Library of Medicine for inclusion in the UMLS. The NCI Director's Advisory Group and the Vocabulary Executive Group should determine other opportunities that the Institute should explore to get new cancer-related concepts and terminology into use across the full range of medical vocabulary and clinical practice.

Long-Term Goals

When the NCI governance structure is mature and its involvement in standards and vocabulary development and utilization is well established, emphasis should shift to helping the cancer community benefit from these activities. It will take several years to develop adequate communication about information exchange opportunities and requirements and the standards and vocabulary needed to support them.

NCI is working on improvements to its public websites like PDQ, hoping to make it easier for individual patients to find appropriate clinical trials. These improvements are possible because of improved information sharing among NCI systems, supporting both protocol development and information sites. Adoption of standards and vocabulary conventions underpin such improved services to the community. However, many in the cancer community are uninformed, or are inadequately informed, about the role of standards and controlled vocabulary in facilitating modern information sharing and utilization. The NCI must help increase understanding of these issues, and the NCI Director's Advisory Group could play an important role in identifying ways to spread the word.

Much of the progress being made today in understanding cancer, its treatment, and its prevention comes from molecular and genetic research and from moving basic science insights rapidly into clinical intervention or recommendations for prevention. The terminology used by molecular and genetic science is unfamiliar to many in the cancer community. Some of these terms will be available in NCI Thesaurus before they are modeled in vocabularies with a more general focus, and arrangements are underway to make them available to the larger community. In the long term, however, the NCI may have to go beyond vocabulary modeling into ontology development.

The key to making clear the relevance of molecular or genetic concepts to clinical practice or to cancer prevention is through semantic relationships that link them. The description logic used in the NCI Thesaurus appears to be capable of representing a useful amount of such semantic information. Still, workable rules to use the capability while avoiding the well-known pitfalls of modeling are yet to be developed. Developing satisfactory rules for such semantic modeling is a difficult long-term challenge for NCI to address.

Conclusion

Shared information can enable new science through collaboration, clinical excellence, lowered costs, better programmatic analyses, and research efficiency. This vision depends on sharing vocabulary and other intellectual property that costs a great deal to create. When the government does not defray the cost of development, the producer must charge for use of the property to cover the costs of development. Commercial considerations, combined with the highly unsettled state of intellectual property law in the digital arena, may delay or derail some of the information sharing that would most benefit the community.

The cancer community will need to find a voice in the ongoing political and legal dialogue about intellectual property. Organizations wishing to merge public domain content into proprietary offerings seem to have concerns about losing control of their intellectual property, while organizations wishing to use intellectual property seem to be uncertain of the boundaries of fair use. Across the biomedical marketplace, the pricing of such intangible products as sequences and vocabulary is so inconsistent and confusing that it is impeding adoption of useful products and techniques.

The investment in standards and vocabulary described in this chapter will undoubtedly prove beneficial to the NCI and to the larger community, making accurate information available where it can do the most good. To reap the potential benefits, the Institute must rise to the challenge of addressing difficult technical and process issues. At the same time, it must pay special attention to legal and economic issues, which will determine the scope of the benefits cancer care can realize.

References

Brown, S. 2000. *Veterans Administration*, personal communication (NO MORE INFO).

Campbell, K. E. 1997. "Distributed Development of a Logic-Based Controlled Medical Terminology." Dissertation submitted to Stanford University. *Dissertation Abstracts International,* 01 599 569 (We don't believe there is an issue or volume number – the abstract number identifies it).

Chute, C. and Langlotz, C., Chairs. 2000. *Translating Cancer Research into Cancer Care*. Final Report of the Long Range Planning Committee, National Cancer Institute. (This will be available through a website that is in development. We will send the NCI EVS website sometime in December.)

Gray, P. and Spaholtz, B., Chairs. 2000. *Goals and Principles for Establishing an Integrated NCI-Wide Coding System*. Report of the Institute Information Systems Advisory Group Subcommittee on Coding, National Cancer Institute (It is not a public document at this time. It is only available on the NCI Intranet. Perhaps it should be annotated as draft?).

Harris, N.L., Jaffe, E.S., Diebold, J., Flandrin, G, Muller-Hermelink, H.K., Vardiman, J., Lister, T.A., Bloomfield, C.D. "World Health Organization Classification of Neoplastic Diseases of the Hematopoietic and Lymphoid Tissues: Report of the Clinical Advisory

Committee Meeting – Airlie House, Virginia, November 1997." *Journal of Clinical Oncology*, 17: 3835-3849, 1999.

Hochhalter, B. May 2000. "Putting Terminology in the Business Plan: Using SNOMED/RT in Kaiser Permanente Information System." Paper given at *Towards an Electronic Patient Record (TEPR 2000), San Francisco, CA.*

Klausner, R. 1999. *The Nation's Investment in Cancer Research, a Budget Proposal for Fiscal Year 2001*. National Cancer Institute, p. 52. (http://2001.cancer.gov)

Makower D [a]. Rozenblit A. Edelman M. Augenlicht L. Kaufman H. Haynes H. Zwiebel J. Wadler S. Oncolytic viral therapy in hepatobiliary tumors. *Cancer Investigation. 18(SUPPL. 1). 2000. 111-112.*

Gilfillan, L, Haddock, G, Borek, S. "Knowledge Management Across Domains", *SPIE*, in press, 2001. (http://www.theresearchplace.com/SPIEPaper.htm)

Readings

For more information on the new commercial version of The Science Place, see http://www.TheResearchPlace.com.

College of American Pathologists, 2000. *SNOMED/RT Documentation (Draft).*

College of American Pathologists, 2000. *SNOMED Clinical Terms Technical Specification (Draft).*